Data 100, Discussion 8

Suraj Rampure

Wednesday, October 16th, 2019

Agenda

- Motivating linear regression
- Correlation
- Bootstrapping

Lots of demos. As per usual, everything will be posted at

http://surajrampure.com/teaching/ds100.html

Review – Summary Statistics

Before: we considered a collection of data points $\{x_1, x_2, \ldots, x_n\}$, and we wanted to come up with a **summary statistic** c for this data, that is the "best", in some sense.

We defined our **loss** for a single point in terms of the prediction error, $x_i - c$. We often used the L_2 loss, and we will continue doing that now.

actual − pred

L_2 loss for a single point: $(x_i - c)^2$

Average L_2 loss for entire dataset:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2}_{\text{empirical risk}}$$

Simple Linear Regression

Now, suppose we have a collection of data points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Instead of creating a summary statistic for x or y individually, we want to **model y as a linear function of x**, i.e.

$$\hat{y}_i = \beta x_i$$

$y = mx$

or, if we'd like to include an intercept term,

$$\hat{y}_i = \beta_1 x_i + \beta_0$$

$y = mx + b$

$$\frac{1}{n} \sum (\text{actual} - \text{pred})^2$$

$$= \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

$$= \frac{1}{n} \sum (y_i - \beta_1 x_i - \beta_0)^2$$

empirical risk

Ordinary Least Squares

Suppose we're given $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and want to fit a linear model $y = \beta_1 x + \beta_0$, using MSE (i.e. L2) loss.

Our objective function is

$$L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

One way to solve: Take partial derivatives with respect to β_0, β_1. Solve for β_0 and β_1.

$\frac{\partial L}{\partial \beta_0} = 0$, $\frac{\partial L}{\partial \beta_1} = 0 \rightarrow$ 2 equations, 2 unknowns, solve

income = β_1(height) + β_2(shoe size) + β_3(GPA) + β_0

too many derivatives!!!

actual - pred : residue

Let's try and rewrite this in vector form.

$$\boxed{L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2}$$

$$X\beta = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix}$$

We can say the following:

$\beta = \begin{bmatrix}\beta_0 & \beta_1\end{bmatrix}^T$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\boxed{L(\beta) = \frac{1}{n}\|y - X\beta\|_2^2}$$

$$\Uparrow$$

$y = \begin{bmatrix}y_1 & y_2 & \dots & y_n\end{bmatrix}^T$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Correlation

The concept of correlation is intimately tied to the idea of simple linear regression.

$$r(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_x}{SD(x)} \right) \left(\frac{y_i - \mu_y}{SD(y)} \right)$$

r, denoted the **correlation coefficient**, is a value between -1 and 1.

- A value of 0 denotes absolutely no linear correlation.
- As r approaches 1 (or -1), the strength of the correlation between x and y increases.
- The sign of r tells us whether our correlation is positive (up and to the right) or negative (down and to the right)

$r = -1$:

$r = 1$:

Bootstrapping

Refer here for my slides from Data 8 on bootstrapping.

1. Obtain a sample from the population of interest. Compute the sample statistic $\hat{\theta}$.
2. Repeatedly sample (with replacement!) from our obtained sample.
3. For each bootstrap sample, compute a sample statistic. Generate $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_{10000}$.
4. Look at the distribution of all bootstrapped sample statistics, and see where the original sample statistic lies.

- In Data 8, we primarily bootstrapped to create a confidence interval for some population parameter, e.g. the mean of the heights of students at Berkeley.
- Towards the end of Data 8, and now, we will instead bootstrap to create a confidence interval for the slope of a linear relationship, i.e. for α in $y_i = \alpha x_i$